

APPENDIX F

Input: 1. A continuous OR categorical dataset X

2. Target variable y is continuous

Output: A filtered continuous or categorical dataset

Process:

1. Bin the target y into a categorical variable bin_y

2. Calculate correlation of each variable x with y.

If x is a continuous variable, the correlation is Pearson's R between x and y; If x is a categorical variable, the correlation is Cramer's V between x and bin_y.

3. Let n equal the number of variables in the input dataset and k is the number of variables to be kept.

If (n <= 50) k = n ;

Else If n <= 100 k = 50 + round(0.7 * (n - 50)) ;

Else If n <= 200 k = 85 + round(0.5 * (n - 100)) ;

Else k = 135 + round(0.3 * (n - 200)) ;

End If

4. Sort the variables based on the absolute correlation value in descending order, and keep the first k variables. Store their indexes and correlation values with y.